Adversarial Machine Learning in Recommender Systems (AML-RecSys)

Yashar Deldjoo
Polytechnic University of Bari
Bari, Italy
yashar.deldjoo@poliba.it

Tommaso Di Noia
Polytechnic University of Bari
Bari, Italy
tommaso.dinoia@poliba.it

Felice Antonio Merra*
Polytechnic University of Bari
Bari, Italy
felice.merra@poliba.it

ABSTRACT

Recommender systems (RS) are an integral part of many online services aiming to provide an enhanced user-oriented experience. Machine learning (ML) models are nowadays broadly adopted in modern state-of-the-art approaches to recommendation, which are typically trained to maximize a user-centred utility (e.g., user satisfaction) or a business-oriented one (e.g., profitability or sales increase). They work under the main assumption that users' historical feedback can serve as proper ground-truth for model training and evaluation. However, driven by the success in the ML community, recent advances show that state-of-the-art recommendation approaches such as matrix factorization (MF) models or the ones based on deep neural networks can be vulnerable to adversarial perturbations applied on the input data. These adversarial samples can impede the ability for training high-quality MF models and can put the driven success of these approaches at high risk. As a result, there is a new paradigm of secure training for RS that takes into account the presence of adversarial samples into the recommendation process.

We present *adversarial machine learning in Recommender Systems* (AML-RecSys), which concerns the study of effective ML techniques in RS to fight against an adversarial component. AML-RecSys has been proposed in two main fashions within the RS literature: (i) *adversarial regularization*, which attempts to combat against adversarial perturbation added to input data or model parameters of a RS and, (ii) *generative adversarial network* (GAN)-based models, which adopt a generative process to train powerful ML models. We discuss a theoretical framework to unify the two above models, which is performed via a *minimax game* between an adversarial component and a discriminator. Furthermore, we explore various examples illustrating the successful application of AML to solve various RS tasks. Finally, we present a global taxonomy/overview of the academic literature based on several identified dimensions, namely (i) research goals and challenges, (ii) application domains and (iii) technical overview.

*Authors are listed in alphabetical order. Corresponding author: Felice Antonio Merra (felice.merra@poliba.it).

WSDM '20, February 3–7, 2020, Houston, TX, USA

ACM ISBN 978-1-4503-6822-3/20/02.
https://doi.org/10.1145/3336191.3371877

ACM Reference Format:
Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2020. Adversarial Machine Learning in Recommender Systems (AML-RecSys). In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20), February 3–7, 2020, Houston, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3336191.3371877

1 INTRODUCTION AND CONTEXT

In the frenetic times of the 21st century, where users are facing a new form of information explosion that is exhausting the Web-space, recommender systems have emerged as a principal defence strategy against consumer over-choice. Recommender systems have drastically facilitated the user experience (with lots of services) by embracing user-centred design best practices as they provide an interactive experience and personalize the informational flows for each user independently but also taking into consideration the collaborative behaviour of other users.

Different recommendation models use as input explicit or implicit user feedback and/or item features. Based on the available information, three classes of recommendation models can be defined: collaborative filtering (CF), content-based filtering (CBF) models and hybrid ones [27]. CF builds on the fundamental assumption that personal tastes correlate and users who expressed similar taste in the past tend to agree in future as well [13]. In order to compute predictions, CF approaches to measure the appeal of each unknown item for a given user by analyzing the correlations among the behavioural data of users, which can be implicit (their previous clicks, check-ins) or explicit (their rating scores). On the other hand, CBF [19] relies only on the interactions of single users together with item content features/attributes to create a model of users (aka user profiles) that represents the nature of their interests. Among the main paradigm of recommendation, collaborative filtering (CF) models have been the practical choice of industrial engines and the subject for highest academic research, due to their superb recommendation quality compared with content-based filtering (CBF) models, when sufficient amount of interaction data are available.

We find it insightful to provide a historical development of CF models, which can be divided into several overlapping research lines described below and illustrated in Figure 1.

Classical non-neural CF era: the beginning of this era dates back to the 1990s and it is still ongoing. The research in this era is characterized as the one in which research on RS is mainly focused on improving the utility (usefulness) of recommendations commonly measured in terms of either accuracy of estimating ratings (e.g., evaluated in terms of RMSE, MAE) or accuracy of estimating rankings (a.k.a. top-N recommendation accuracy). Almost three decades on CF research, have resulted on a rich set of CF models

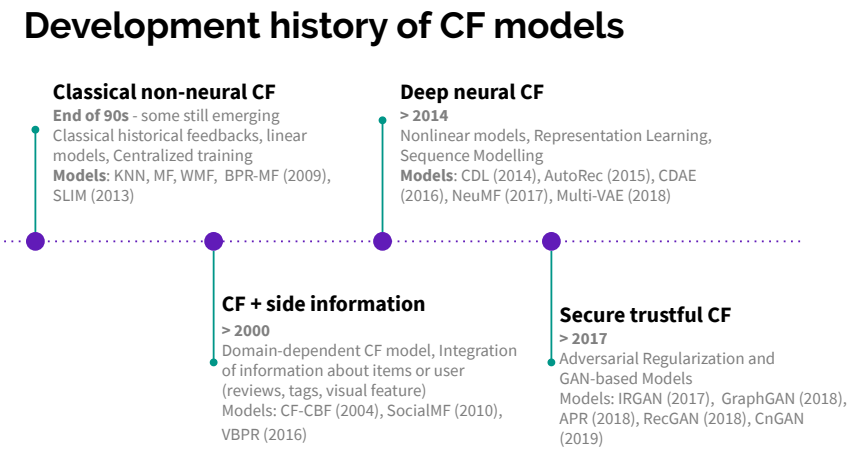

Figure 1: Development history of modern approaches to recommendation, which are mostly centered around CF model.

(and evaluation metrics), which can be roughly placed into one of the categories of *memory-based* such a nearest neighborhood approaches, and *linear model-based* based on matrix factorization (MF) latent factor models [11].

In the course of time, other measures such as diversity and novelty of recommendations became important. Therefore in early the 2000s, part of research in the RS community was focused on improving the *beyond-accuracy measures* including (but not limited to) novelty, diversity and coverage. Furthermore, several research works were focused on *stability and robustness* of CF models in the presence of attacks such as fake ratings. These type of *shilling attacks* were **hand-engineered** and were usually designed to alter true user rating distribution for a malicious purpose or profit-driven motivations such as market penetration [10, 16].

CF + side information era: "Recommendation is not a one-size-fits-all problem" [11]. In the early 2000s, the quest for designing *domain-dependant* recommendation models was on the rise. Part of this owed to the fact that the start of 21st century was coinciding with the period in which mobile/smartphones, PDAs and tablets were ubiquitously available and they were, in fact, responsible for the generation of torrents of digital data, such as images, text, audio, videos. All these sources of information motivated a new generation of hybrid CF models, which could use the information beyond the user interaction, known as *CF models using side information.*

A large body of works has been published on such hybrid CF models. They can be classified based on if these model resort to side information related to users and items, such as social networks [3], user-contributed information (tags, metadata reviews) [24], multimedia content such image and audio signal [9] and the information related to the interaction between users and item, e.g., time [1].[1] Regardless of the approach and domain, the main concentration of the research in this period was on improving recommendation accuracy and beyond-accuracy measures.

[1]We can see context-aware recommendation models, a child of this era.

Deep neural CF era: In the most recent years, recommendation models based on deep learning (DL) or "neural" technology have become predominately important. Deep neural networks (DNN) received a hype after the convolutional neural network (CNN) architecture named AlexNet proposed by Krizhevskyet et. al. reached a ground-breaking accuracy for image classification in 2012 [22]. A large part of this success owes in the availability of large labelled datasets and computational resources nowadays, which allows the system to learn a massive amount of model parameters and ultimate high-quality features [23].

Although DNNs have been used to solve a variety of tasks in RS domain, such as feature extraction (e.g., via using CNNs), sequence and temporal modelling (e.g., via using RNNs), here we keep our attention on the CF preference predicting model. In this line, a growing number of research works have used DNNs to parameterize MF with neural architectures, which in turn enables them to model the interaction between users' and items' latent factors in a non-linear fashion thus giving them the capability to model/solve more complex inference problems [18, 34]. Again, the success in RS evaluation here has been evaluated based on a dominant paradigm based on accuracy and/or beyond-accuracy aspects depending on the application in question.

Secure trustful CF era: In [15], the authors demonstrate that by adding small adversarial perturbations to an image of pandas, they can fool a well-trained classifier to misclassify the image as a gibbon with high confidence and the impact of perturbations added is barely observable with naked eyes. As DNN became powerful to solve many inference problems in various domains including ML and RS, *security* of the models became more critical.

The main objective of AML-RecSys is to study effective ML architectures and training procedures in the design space of RS with the introduction of an adversarial component. If we look at attack scenarios, what makes AML-RecSys different from shilling attacks identified in the classical *non-neural CF model era* is that shilling attacks are hand-engineered fake user profiles whereas AML-RecSys deals with an **automated** process to generate perturbed profiles

or content. AML-RecSys is a reminder of the metaphoric proverbial: "Security is sometimes thought of like a chess game between two players. For a player to win, it is not only necessary to have an effective strategy, but one must also anticipate the opponent's response to that strategy" [20].

Although the research in AML-RecSys is characterized by aiming to improve the robustness and stability of recommendation models, we will show that it can be used to improve the general performance of recommendation toward improved accuracy and various other objectives among others including improved beyond-accuracy recommendation capability, improved dynamic negative sampling in pairwise learning or complementary fashion outfit recommendation (e.g., recommending pair of jeans that match a given shirt).

2 OUTLINE OF THE TUTORIAL

In this section we present the proposed tutorial schedule.

Introduction to Recommender Systems (20 mins) We start the tutorial with a general introduction to recommender systems. RSs support user's decision making process by pointing them to a small set of items (top-N list [2]) out of a large catalogue [36]. We present typical models of RS as collaborative filtering (CF), content-based filtering (CBF) and hybrid thereof [27].

Deep Learning in RS: (20 mins) We give a brief introduction on how the DL field has been exploited to solve recommendation tasks. For instance, collaborative deep learning (CDL) [31], is a deep CF model that completes the user-item feedback matrix by learning a deep representation of users-items relations jointly with the extraction of deep latent features from items' side information; network-based collaborative filtering (NCF) [18] learns user-item interactions representation directly through a multi-layer perceptron, and collaborative denoising auto-encoders (CDAE) [34] is a state-of-the-art auto-encoder-based approach that reconstructs the user-feedback sparse profile in the output layer by integrating also users, and items, side information. Other intriguing and state-of-the-art applications of DL in recommendation domain are convolutional neural networks (CNNs) for *automatic feature extraction* (e.g., audio and visual features [7, 8]), and recurrent neural networks (RNNs) for *sequence modeling* of user preferences [26].

Adversarial Machine Learning in RS: (115 mins) AML is a field of machine learning which has been introduced to demonstrate the possibility of failure for current deep models under adversarial-perturbed examples [29]. We present an in-depth analysis of how AML techniques have been applied to solve a variety of recommendation task, covering the following topics:

(1) *Notions of AML.* We present here basic definitions of AML of adversarial examples, adversarial perturbations, adversaries, adversarial training and the description of GAN [14].

(2) *Foundations of AML-RecSys.* We present here a categorization of the two AML application in recommendation scenarios: the *adversarial regularization* and the *GAN-based recommendation.* In particular, we examine and present two pioneer works for each of AML-RecSys. Adversarial Personalized Learning (**APL**) [17] verifies the effectiveness of adversarial perturbations against recommender models (i.e., matrix factorization) and applies *adversarial regularization* techniques to make the system more robust to adversarial noise. This work is at the core of different novel works on recommender models [30, 35] Information Retrieval-GAN (**IRGAN**) [32] is the first attempt to unifies generative and discriminative information retrieval models under a unique framework. They use the proposed GAN-based framework into item recommendation task with surprising results by outperforming baseline recommender models (r.g., BPR and LambdaFM) on both precision and ranking-based performances. The proposed approach has inspired a lot of GAN-based recommender models [4, 6]

(3) *Categorization based on the goals.* We identify two main adversarial goals in recommendation scenario: (a) improve robustness of recommender model and (b) improve the utility of recommender system. The first goal focuses on applications of AML to attack RS to deteriorate the recommendation performance and protect the system. This goal can be modelled as a *minimax game* played between an adversary, who crafts adversarial examples in order to alter RS performance, and the recommender engine, that is trained with different techniques to became robust to such noise. The second goal is based on the application of AML to improve the utility of recommendation performances. For instance, the *minimax-game* is applied to improve the performance of RecSys by generating more informative negative samples, instead of randomly chosen ones, in order to learn better users and items latent vectors in learning-to-rank tasks [12, 33], as well as to fit the generator of a GAN-based architecture in predicting missing scores (e.g., ratings, implicit feedback) by leveraging either user-item side content [6, 32] or contextual information [4, 37]. There are also GAN-based solutions to augment training dataset [5] in order to generate data that follows the real data distribution.

(4) *Application Domains.* AML-RecSys has been applied in different domains. We have classified them into four main groups: Media & Entertainment, E-Products, Social-Computing, and Information-based domains. It is interesting to observe the use of the transferability and generative properties of GAN-based solutions in the case we want to recommend complementary items [21, 28], and cross-domain recommendation [12, 25].

(5) *Technical Analysis.* We present technical and architectural solutions in recommender systems looking at the application of adversarial learning. Particularly, we structure the discussion by considering a classification in GAN-based and Adversarial-learning-based solutions.

Conclusions, Grand Challenges and Discussion (25 mins) We round off the tutorial by giving a brief summary, communicating the main take away messages, and providing some practical guidelines for researchers new to the area of AML-RecSys. Via the identification and discussion of the grand challenges, we further guide researchers new to the topic, and hopefully, help them shape their ideas for future research directions on this interesting field. Such challenges include, among others: (i) the development of approaches to identify possible security issues of deep learning-based recommendation models, in particular in the light of recent advances in adversarial machine learning, (ii) propose novel defence approaches to improve the robustness of the recommendation

system and (iii) promote the research towards novel recommendation models that can exploit the advances of AML verified in the computer vision domain.

3 SUPPORT MATERIALS.

The tutorial is supported by a GitHub page with an overview of the program, *tutorial slides*, all the references and presenters details[2].

REFERENCES

[1] Vito Walter Anelli, Vito Bellini, Tommaso Di Noia, Wanda La Bruna, Paolo Tomeo, and Eugenio Di Sciascio. 2017. An Analysis on Time- and Session-aware Diversification in Recommender Systems. , 270–274 pages.

[2] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Azzurra Ragone, and Joseph Trotta. 2019. Local Popularity and Time in top-N Recommendation. In *ECIR (1) (Lecture Notes in Computer Science)*, Vol. 11437. Springer, 861–868.

[3] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*. 635–644.

[4] Homanga Bharadhwaj, Homin Park, and Brian Y. Lim. 2018. RecGAN: recurrent generative adversarial networks for recommendation systems. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. 372–376.

[5] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jaeho Choi. 2019. Rating Augmentation with Generative Adversarial Networks towards Accurate Collaborative Filtering. In *WWW*. ACM, 2616–2622.

[6] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jung-Tae Lee. 2018. CFGAN: A Generic Collaborative Filtering Framework based on Generative Adversarial Networks. In *CIKM*. ACM, 137–146.

[7] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2018. Audio-Visual Encoding of Multimedia Content to Enhance Movie Recommendations. In *Proceedings of the Twelfth ACM Conference on Recommender Systems*. ACM.

[8] Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*. 450–455.

[9] Yashar Deldjoo, Maurizio Ferrari Dacrema, Mihai Gabriel Constantin, Hamid Eghbal-zadeh, Stefano Cereda, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2019. Movie genome: alleviating new item cold start in movie recommendation. *User Model. User-Adapt. Interact.* 29, 2 (2019), 291–343.

[10] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2019. Assessing the Impact of a User-Item Collaborative Attack on Class of Users. *ImpactRS@RecSys'19 Workshop on the Impact of Recommender Systems* (2019).

[11] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. 2011. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction* 4, 2 (2011), 81–173.

[12] Wenqi Fan, Tyler Derr, Yao Ma, Jianping Wang, Jiliang Tang, and Qing Li. 2019. Deep Adversarial Social Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. 1351–1357.

[13] David Goldberg, David A. Nichols, Brian M. Oki, and Douglas B. Terry. 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM* 35, 12 (1992), 61–70.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2672–2680.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[16] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Polat. 2014. Shilling attacks against recommender systems: a comprehensive survey. *Artif. Intell. Rev.* 42, 4 (2014), 767–799.

[17] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. 355–364.

[18] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[19] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. 263–272.

[20] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar. 2011. Adversarial machine learning. In *AISec*. ACM, 43–58.

[21] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. 2017. Visually-Aware Fashion Recommendation and Design with Generative Image Models. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*. 207–216.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[23] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikainen. 2018. A survey of recent advances in texture representation. *arXiv preprint arXiv:1801.10324* (2018).

[24] Xia Ning and George Karypis. 2012. Sparse linear methods with side information for top-n recommendations. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*. 155–162.

[25] Dilruk Perera and Roger Zimmermann. 2019. CnGAN: Generative Adversarial Networks for Cross-network user preference generation for non-overlapped users. In *WWW*. ACM, 3144–3150.

[26] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *ACM Comput. Surv.* 51, 4 (2018), 66:1–66:36.

[27] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer, 1–34.

[28] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. 2018. Compatibility Family Learning for Item Recommendation and Generation. In *AAAI*. AAAI Press, 2403–2410.

[29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.

[30] J. Tang, X. Du, X. He, F. Yuan, Q. Tian, and T. Chua. 2019. Adversarial Training Towards Robust Multimedia Recommender System. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1.

[31] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative Deep Learning for Recommender Systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. 1235–1244.

[32] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 515–524.

[33] Qinyong Wang, Hongzhi Yin, Zhiting Hu, Defu Lian, Hao Wang, and Zi Huang. 2018. Neural Memory Streaming Recommender Networks with Adversarial Training. In *KDD*. ACM, 2467–2475.

[34] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. 2016. Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*. 153–162.

[35] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial Collaborative Neural Network for Robust Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. 1065–1068.

[36] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1 (2019), 5:1–5:38.

[37] Wei Zhao, Benyou Wang, Jianbo Ye, Yongqiang Gao, Min Yang, and Xiaojun Chen. 2018. PLASTIC: Prioritize Long and Short-term Information in Top-n Recommendation using Adversarial Training. In *IJCAI*. ijcai.org, 3676–3682.

[2] https://github.com/sisinflab/amlrecsys-tutorial